PROCEEDINGS OF SPIE

SPIEDigitalLibrary.org/conference-proceedings-of-spie

Signaling local non-credibility in an automatic segmentation pipeline

Joshua H. Levy, Robert E. Broadhurst, Surajit Ray, Edward L. Chaney, Stephen M. Pizer

Joshua H. Levy, Robert E. Broadhurst, Surajit Ray, Edward L. Chaney, Stephen M. Pizer, "Signaling local non-credibility in an automatic segmentation pipeline," Proc. SPIE 6512, Medical Imaging 2007: Image Processing, 65123Q (26 March 2007); doi: 10.1117/12.709015

SPIE.

Event: Medical Imaging, 2007, San Diego, CA, United States

Honorable Mention Poster Award

Signaling Local Non-credibility in an Automatic Segmentation Pipeline

Joshua H. Levy[a], Robert E. Broadhurst[a], Surajit Ray[b], Edward L. Chaney[c] and Stephen M. Pizer[a]

Departments of Computer Science[a] and Radiation Oncology[c]
University of North Carolina, Chapel Hill, NC, USA;
[b]Department of Mathematics and Statistics, Boston University, Boston, MA, USA

ABSTRACT

The advancing technology for automatic segmentation of medical images should be accompanied by techniques to inform the user of the local credibility of results. To the extent that this technology produces clinically acceptable segmentations for a significant fraction of cases, there is a risk that the clinician will assume every result is acceptable. In the less frequent case where segmentation fails, we are concerned that unless the user is alerted by the computer, she would still put the result to clinical use. By alerting the user to the location of a likely segmentation failure, we allow her to apply limited validation and editing resources where they are most needed.

We propose an automated method to signal suspected non-credible regions of the segmentation, triggered by statistical outliers of the local image match function. We apply this test to m-rep segmentations of the bladder and prostate in CT images using a local image match computed by PCA on regional intensity quantile functions.

We validate these results by correlating the non-credible regions with regions that have surface distance greater than 5.5mm to a reference segmentation for the bladder. A 6mm surface distance was used to validate the prostate results. Varying the outlier threshold level produced a receiver operating characteristic with area under the curve of 0.89 for the bladder and 0.92 for the prostate. Based on this preliminary result, our method has been able to predict local segmentation failures and shows potential for validation in an automatic segmentation pipeline.

Keywords: Validation

1. INTRODUCTION

An automated process for segmenting 3D medical images allows clinicians to solve previously intractable problems. One such application is segmenting a large collection of historical images of patients in order to correlate the treatment a patient received with his clinical outcome. Limits on human time and attention restrict the number of images that can be processed with costs that scale linearly with the number of images to be analyzed and the number of relevant slices per image. However, an automated system can be scaled and distributed in order to segment an arbitrarily large data set. A second application where automated segmentation can solve a previously intractable problem is related to images when the underlying data changes rapidly. During the time a careful human rater spends segmenting the structures of interest in the image, the data changes to the point where the image is no longer an accurate portrait of the system, and the segmentation of image data is irrelevant to the medical problem. An automated system may be able to provide a sufficiently rapid response so that it produces a segmentation during the short interval during which the image can be used.

An automated system that produces clinically acceptable segmentations allows progress to be made on these problems. However such a system brings with it a jeopardy that a segmentation failure will go undetected and will be put to use. For the same reasons of cost that manual segmentations cannot be performed in these situations, fully manual validation of automatic segmentations is not a viable option. In the situation where the underlying biology is rapidly changing, the validity of the image would expire before the validity of the image

Send correspondence to Joshua H. Levy: levy@cs.unc.edu

Medical Imaging 2007: Image Processing, edited by Josien P. W. Pluim, Joseph M. Reinhardt,
Proc. of SPIE Vol. 6512, 65123Q, (2007) · 1605-7422/07/$18 · doi: 10.1117/12.709015

segmentation could be established. Because of these limitations, we propose an automated method for rapidly detecting and signaling the location of non-credible regions on a segmentation surface.

Our method for identifying non-credible regions requires a geometry to image match ("image match") function that can be evaluated region by region at the segmentation surface. Let I denote an image and let m denote some parameterization of the segmentation of the structure of interest in the image data. The image match $f(m, I)$ serves as a proxy for the distance from the segmentation surface to the unknown surface representing the ground truth for the object in the image. We require that I can be decomposed into object-relative regions $\{x\}$ and that $f(m, I)$ can be decomposed into independent terms $f_x(m, I)$ localized to each region. Each of these local image match terms indicates the goodness of fit for a subregion of the segmented object. Statistical analysis on the distribution of the value of each local image match term over a set of well fit training cases produces tests to identify non-credible regions on a segmentation surface.

In regions where the local image match has an unusually poor value, we expect there will be a large error to the unknown truth, so the segmentation should be considered non-credible. We assume, without loss of generality, that the value of the local image match function increases as the localized goodness of fit decreases. Given a new image $\hat{I}$ and its segmentation $\hat{m}$ we want to identify the set of regions:

$$X^* \quad = \quad \left\{ x : p\left[f_x\left(m, I\right) \geq f_x\left(\hat{m}, \hat{I}\right)\right] < \rho \right\} \tag{1}$$

for some critical value ρ. These are the non-credible region of the segmentation surface.

The test, (1), relies on our ability to learn the probability distribution for the local image match terms. Certain classes of local image match functions allow us to make principled assumptions about this distribution. When $f_x(\cdot, \cdot)$ is the log-likelihood of a multivariate Gaussian, its value is a Mahalanobis distance and thus is a χ^2 random variable. This is the case when the image match function is produced via Principal Component Analysis (PCA) on a training set. The image match invented by Broadhurst[1] is trained in this manner, and we have used it in the work presented here.

Given our ability to detect non-credible regions using (1), we must now decide how our system should proceed when such a region is detected. Three options for handling this situation are:

1. Notify the user and defer to the user's judgment on how to react
2. Notify the user and force the user to interact with a segmentation tool
3. Automatically adjust the segmentation.

In this paper we disregard the third option, noting that in the special case of segmentation by optimizing the fit of a deformable shape model[2,3] an optimization that yields an outlier value of the image match function may indicate a segmentation failure but does not prescribe a method to correct it.

We therefore recommend that when a non-credible region is detected the system should notify the user and rely on the user's expert judgment to resolve the situation. In some applications the user may choose to run a semi-automatic segmentation editor, such as the one proposed by Grady[4], in order to correct the segmentation in the non-credible regions. In other applications it may be sufficient for the user to manually verify and correct the segmentation. The benefit of using our method is that manual validation resources are only used in the cases and locations where a segmentation error is likely to have occurred. The prohibitively expensive step of manually verifying each segmentation in its entirety has been eliminated.

2. METHODOLOGY

Our method is to evaluate a localized image match function, region by region, along the surface of a segmentation. Those regions where the image match value is excessively large are considered to have a non-credible segmentation. We prepare a visualization as in Fig. 1 that allows the user to understand the location of suspected segmentation errors. This visualization is analogous to those proposed by Niessen[5] to communicate validation results. Based on the visualization, the user is then responsible for verifying and correcting the segmentation in these regions.

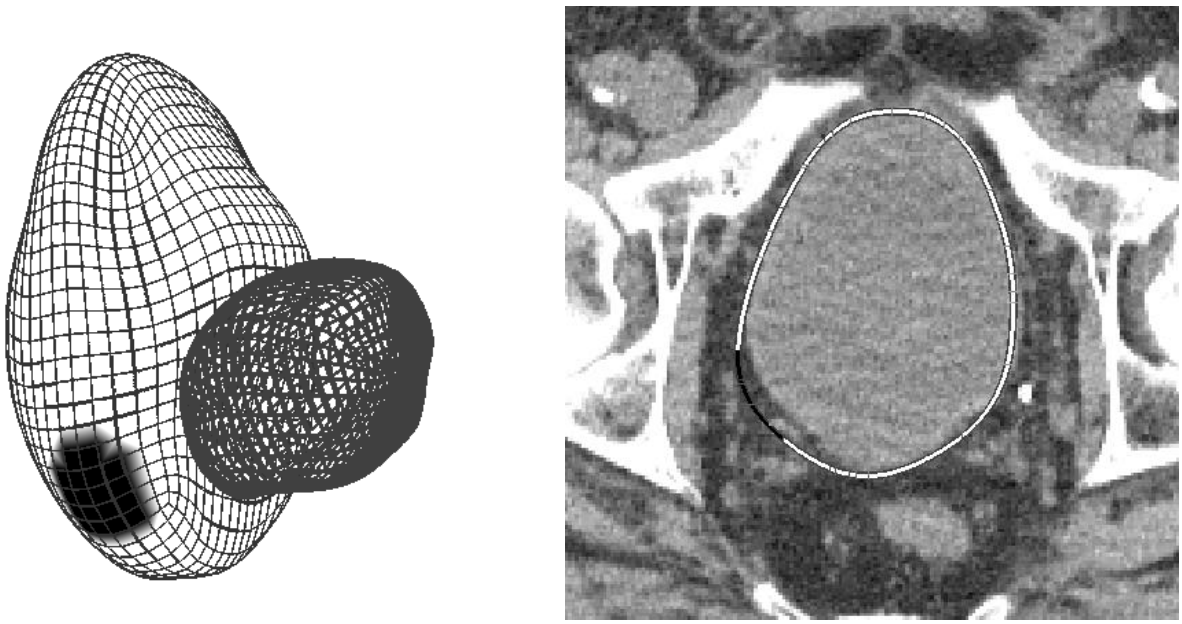

Figure 1. (Left) A display indicating a non-credible region (dark tiles) on a bladder segmentation surface. For orientation purposes, the prostate segmentation surface is shown with unshaded tiles. (Right) The intersection of the same surface with an axial slice of the CT image. In the regions where the segmentation was determined to be credible (white contour) it is well fit to the image data. A significant error is visible in the non-credible (dark contour) region.

In Sec. 2.1 we describe the local image match function that we used to assess the credibility of segmentations of the bladder and prostate in CT images. In Sec. 2.2 we describe the use of the m-rep medial representation to partition a population of objects into corresponding regions.

2.1. Regional intensity quantile functions

The local image match function that we use is based on Broadhurst's regional intensity quantile functions (RIQF).[1] This match requires training on a set of images each of which has a clinically acceptable segmentation. Image intensity histograms $\{H_{x,k}\}$ are recorded for each region x, in each of $\{k\}$ training images. The local image match function measures statistical distance between the observed histogram $\hat{H}_x$ in a region of the target image and the population of histograms observed for the corresponding region of the training set.

Broadhurst has shown that although the histogram space is nonlinear, each histogram H_x can be uniquely mapped onto a quantile function Q_x through the inverse of its cumulative density function (CDF) and that linear statistics are valid on the resulting quantile space. PCA on the training cases yields a mean μ_x and eigenmodes of variation $\{\lambda_{x,i}, v_{x,i}\}$ for the quantile functions. The final local image match value is

$$f_x(m, I) \quad = \quad \sum_i^n \frac{\left((Q_x - \mu_x) \cdot v_{x,i}\right)^2}{\lambda_{x,i}} + \frac{\|r\|^2}{\lambda_r} \tag{2}$$

The first term is the Mahalanobis distance to the intensity quantiles observed in the target case in the PCA space truncated to n eigenmodes of variation. The second term accounts for the residue r outside of this PCA space. r is weighted by the standard deviation $\sqrt{\lambda_r}$ in the training cases that is unaccounted for in the truncated PCA space.

This image match has several properties that are desirable for detecting non-credible regions in image segmentations. We have observed positive correlation between image match value and segmentation quality, suggesting that the outliers of local image match are likely to be regions where a localized failure has occurred. This image

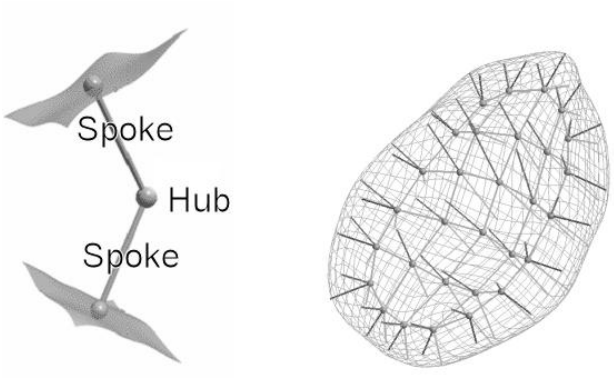

Figure 2. (Left) an example of an m-rep atom. (Right) an m-rep object describing a bladder. Each voxel of the bladder can be identified with m-rep coordinates describing which atom, which spoke of that atom, and how far along the spoke one must travel in order to reach the point.

match models the variability in image intensity for a population of objects and is appropriate for evaluating the segmentation of new image of the same class. Because the image match function is a Mahalanobis distance, its value follows the χ^2 distribution, and critical values on that distribution can be used to establish the threshold ρ for the non-credibility test.

2.2. Surface regions

We use the m-rep medial representation[3] to describe our segmented objects. One advantage of this representation is that it implies an object relative coordinate system that can be used to define corresponding regions across a population of objects. Each m-rep object consists of a grid of sampled medial atoms where each atom is composed of a hub (a location on the medial axis of the object) and a pair of equal length "spoke vectors" which indicates the intersection of the object boundary and its maximal inscribed ball that is centered at the hub. An example of the medial representation of an object is shown in Fig. 2. The object coordinate system maps parameters $\{u, v, \phi\}$ into the volumetric data. Coordinates u and v specify the position of a medial atom in the grid, and coordinate ϕ chooses a spoke at a given hub and specifies how far along the spoke one must travel to reach the point in the volume. The population of m-reps can be constructed so that anatomical correspondences have corresponding coordinates in this object relative coordinate system

We define an object region in the neighborhood of each of the m-rep spoke ends. In each region we record two RIQFs, which are assumed to be statistically independent. One RIQF is recorded from image samples from the interior of the object, and the other from the exterior of the object. Statistics on the RIQFs for a region are learned for a training set of images. The credibility of a segmentation of a new image is assessed by evaluating the local RIQF image match from the training set at the corresponding region in the target object.

3. DETECTING NON-CREDIBLE REGIONS

Our dataset consisted of 80 CT images of the pelvic region of 5 patients receiving adaptive radiotherapy to treat prostate cancer. Each patient's treatment has been fractionated over a series of dates, and a new image was acquired prior to many treatment sessions. The bladder and prostate in each image were segmented by Bayesian optimization over m-rep deformable models.[6] These segmentations required us to train a geometric prior distribution and an image match function.

Let $I_{p,t}$ denote the t^{th} image of patient p and let $m^k_{p,t}$ denote the m-rep segmentation of the bladder ($k = 0$) and the prostate($k = 1$) in $I_{p,t}$. A manual segmentation of each image was performed. These segmentations of patient p on all days other than the target day were used to train the shape prior and the image match. Let $tr^0_{p,t}$ denote a training m-rep fit to the manual segmentation of the bladder in $I_{p,t}$. Each $tr^0_{p,t}$ was constructed so that the m-rep coordinate system preserved anatomical correspondences. The bladder segmentation $m^0_{p,t}$

was produced using a shape prior learned by Principal Geodesic Analysis[7] over the set of m-reps $\left\{tr^0_{p,j} : j \neq t\right\}$, holding p constant, and an image match trained on global RIQFs for the same set of models. The prostate segmentation $m^1_{p,t}$ was produced by applying the same procedure to the prostate training data.

We train local image match functions, for use in the non-credibility test, on regions defined by the m-rep coordinate system. Each region x is defined to be the neighborhood of a sampled m-rep spoke end, which is identified by its object relative coordinates (u, v, ϕ). Because our training m-reps preserve anatomical correspondence, we can train local RIQF statistics on a corresponding region across the training population for an object. This local image match function is used to assess the local credibility of this region in a target case, $m^k_{p,t}$.

The local RIQFs used to train the credibility test for bladder segmentations use three eigenmodes of variation to describe the image intensity quantiles from the exterior of the object and two eigenmodes to describe the interior of the object. We allow an additional two degrees of freedom, one each for the exterior and interior, to account for the residue outside of these eigenmodes. We then make the approximation that $f_x(\cdot,\cdot) \sim {\chi^2}_7$. We can then choose a threshold value f such that $p\left[f_x(\cdot,\cdot) > f\right] < \rho$ for any value ρ by using the known CDF for the χ^2 distribution with seven degrees of freedom. Any region of a bladder segmentation where the image match exceeds f is considered to be non-credible.

We only use the exterior RIQFs to evaluate the local credibility of prostate segmentations because of issues of image contrast between the bladder and prostate. Figure 3 shows two possible segmentations of a prostate on an axial slice of a CT image. One of these segmentations is correct, the other is the result of a gross shift towards the bladder. Because the image intensity patterns for bladder and prostate are similar, the interior histograms for these two segmentations are roughly equivalent, and we would expect them to have similar values of interior image match. Our ability to detect the segmentation error is based on the intensity pattern at exterior of the prostate, in the region away from the bladder. These regional intensity distributions are quite different, and the exterior RIQF image match will be able to distinguish between them. Thus we use exterior RIQFs with three eigenmodes of variability to detect non-credible regions of the prostate. The additional degree of freedom for residue outside of the PCA space allows us to approximate $f_x(\cdot,\cdot) \sim {\chi^2}_4$.

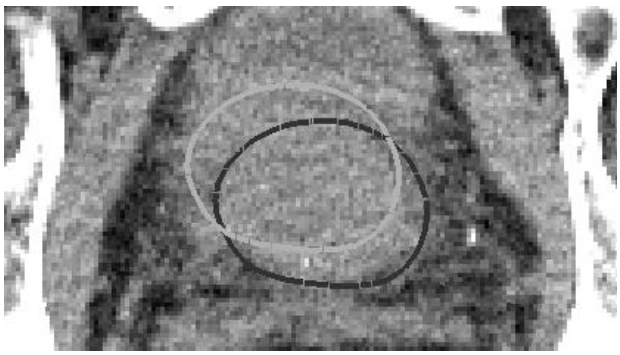

Figure 3. An axial slice of a CT image illustrating why the exterior image match is used to detect non-credibility in prostate segmentations. The two segmentations shown in the image are quite different, yet have similar interior intensity patterns. It is the local exterior intensity pattern that allows us to detect the difference between the acceptable segmentation (dark contour) and the erroneous segmentation (bright contour).

4. VALIDATION

To validate our test for local non-credibility we wish to show that the regions we detected as having outlier values of the local image match function are the same regions where local segmentation failures have occurred. Although we do not know the ground truth for the images in our study, we do have access to manual segmentations for these images. We will then say that a local segmentation failure has occurred when the distance from the representative point for a region (i.e., an m-rep spoke end) to the nearest point on the manual segmentation surface exceeds a limit. That is when $d_x(m, I) > \epsilon$.

The scale of local segmentation errors that we can detect with our test is related to the quality of fits used to train $f_x(\cdot,\cdot)$. Surely the image match cannot be sensitive to errors of the same order of magnitude that were present at training time. For all but two outlier cases of the training m-reps for the bladder, $\max_{x,I} d_x(tr, I) <$ 4mm. For the training m-reps for the prostate, $\max_{x,I} d_x(tr, I) < 4.5$mm. As a result of this measurement we expect our test of non-credibility to be sensitive to segmentation errors where $d_x(m, I) > \epsilon > 4.5$mm. Empirically we found that the test performed well with error defined as $\epsilon = 5.5$mm for the bladder and $\epsilon = 6.0$mm for the prostate.

Using this definition of error, we can validate the test of non-credibility. For an image match threshold of f, the test result at a region on a segmentation surface can be classified in one of four ways:

	$f_x(m, I) \geq f$	$f_x(m, I) < f$
$d_x(m, I) \geq \epsilon$	True-Positive	False-Negative
$d_x(m, I) < \epsilon$	False-Positive	True-Negative

Varying the image threshold f produces a receiver operating characteristic (ROC). An ROC curve, as in Fig. 4 plots the true-positive rate, or *sensitivity* of the test, as a function of the false-positive rate, or $1 -$ *specificity*. This allows us to select an image-match threshold with a clinically acceptable tradeoff between false-positives and false-negatives. The area under the ROC curve (AUC) is an important value. It can be thought of as the probability the test will correctly distinguish between two cases, one that is truly a positive and one that is truly a negative. In our case,

$$AUC \quad = \quad p\left[f_x(m_1, I_1) > f_x(m_2, I_2)\right| \left(d_x(m_1, I_1) \geq \epsilon\right) \wedge \left(d_x(m_2, I_2) < \epsilon\right)] \tag{3}$$

This gives the probability that given two randomly selected segmentation regions, one with a gross error and one that is acceptably close to the reference segmentation, the test will correctly identify the erroneous region.

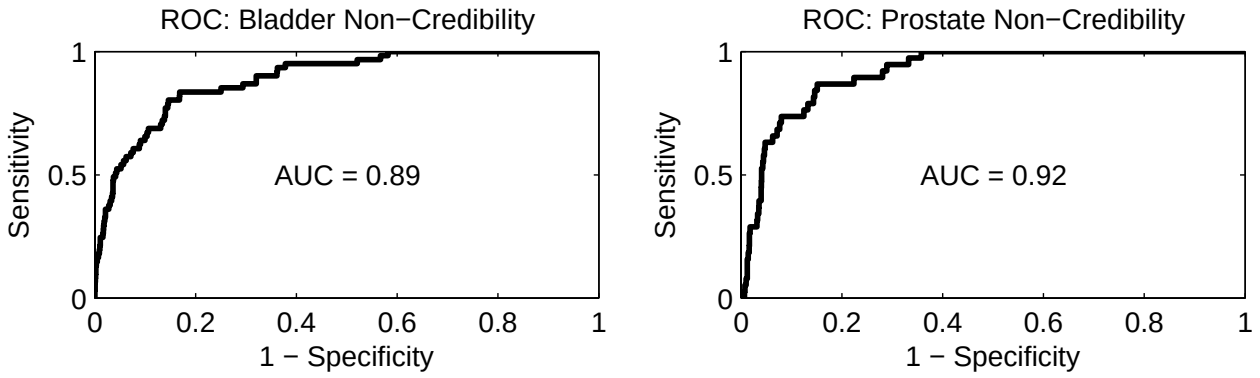

Figure 4. ROC curves characterizing the performance of the test for non-credible regions in bladder and prostate segmentations. For both organs the $AUC \approx 0.9$ indicating that the test successfully distinguishes between segmentation regions that are acceptably close to the reference segmentation and those that are too great of a distance away.

5. RESULTS

The segmentation of the bladders and prostates in 80 CT images were globally quite good. The average over all cases of the average surface distance to the reference segmentation for the bladder was 1.4mm with the maximum over all cases being 2.7mm. For the prostate the average of the average surface distances was 1.2mm with a maximum of 4.0mm.

The distribution of local errors in the bladder segmentations was as follows. In 64 of the 80 segmentations no local error was ever greater than our error definition of 5.5mm. In the remaining 16 cases, 61 out of 1248

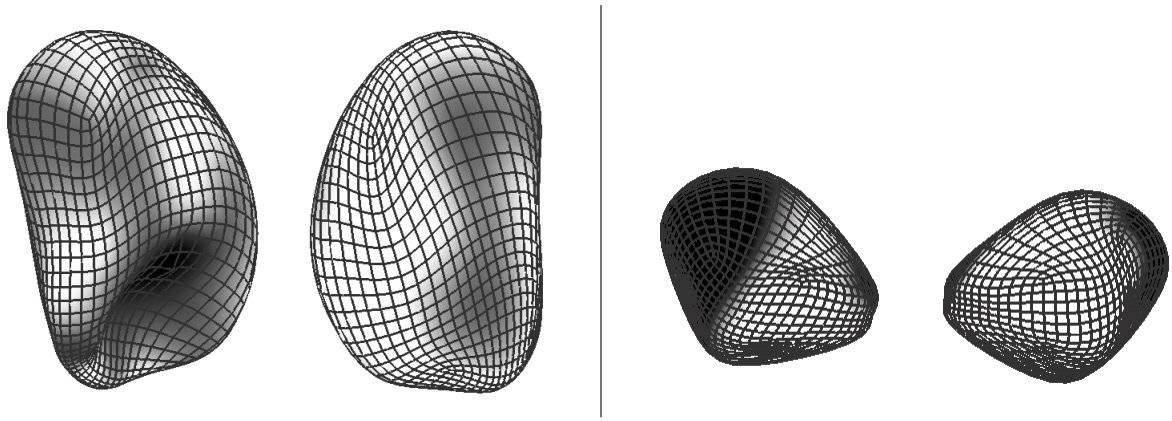

Figure 5. M-rep implied surfaces for a bladder (left) and prostate (right). Darker shading signals regions where errors occurred more frequently.

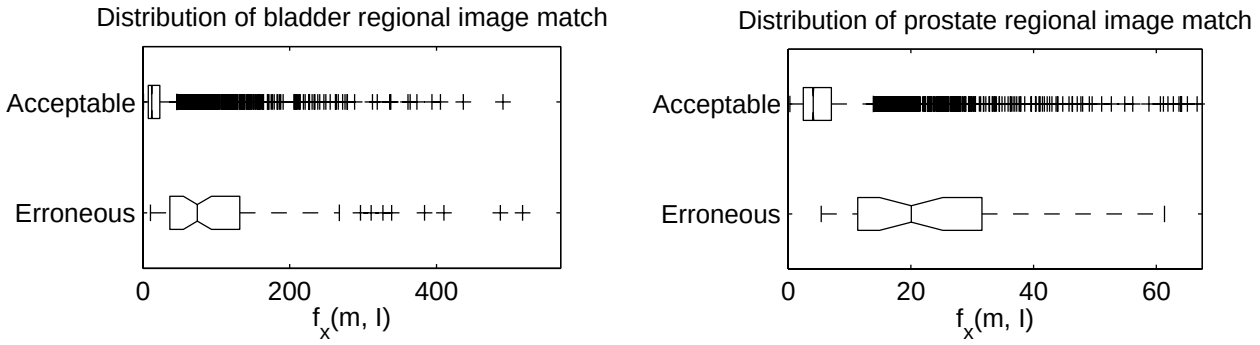

Figure 6. Distribution of the local RIQF image match value for the bladder (left) and the prostate (right) segmentations. The image match values for regions where $d_x(m, I) \geq \epsilon$ tends to be larger than for those regions where $d_x(m, I) < epsilon$

possible regions had a significant local error. The spatial distribution of these regions can be seen in the left pane of Fig. 5. The errors are the most densely concentrated at the low contrast area where the prostate indents into the bladder and on the opposite side of the medial sheet from that location.

There were fewer cases with local errors in the prostate segmentations than there were for the bladder. 76 of the 80 segmentations had no local error greater than our error definition of 6.0mm. In the remaining 4 cases, 38 out of 296 possible regions had a significant local error. The spatial distribution of these regions can be seen in the right pane of Fig. 5. There was a large cluster of errors on the face of the prostate that is adjacent to the bladder. There were also smaller clusters of errors on the opposite faces.

The cases and regions with gross localized errors tend to have larger values of the local RIQF image match function than those that where the local error is below the ϵ threshold. This is presented in Fig. 6. For the median 50% of grossly erroneous bladder regions, $36.76 \leq f_x(m, I) \leq 125.00$. For the median 50% of acceptably segmented bladder regions, $7.19 \leq f_x(m, I) \leq 22.92$. For the median 50% of grossly erroneous prostate regions, $11.35 \leq f_x(m, I) \leq 31.59$. For the median 50% of acceptably segmented prostate regions, $2.48 \leq f_x(m, I) \leq 7.05$.

Figure 4 contains the ROC curve for detecting non-credible regions on bladder and prostate segmentations using our test. AUC is 0.89 for the bladder and 0.92 for the prostate, showing that for nearly 90% of cases the test correctly distinguishes between a region with a significant local error and a region that is correctly segmented. We can use this ROC curve to select the threshold level at which we signal non-credibility. This choice is a tradeoff between the expected true-positive and false-positive rate of the test and needs to be made

with application specific rules in mind. For demonstration purposes we chose a threshold which minimizes the number of false-positive reports.

With this choice of the threshold value, we can now produce the visualizations shown in Fig. 1. In the left pane we see the bladder surface colored so that non-credible regions appear with dark shading. In the right pane we see the intersection of this surface with an axial slice of the CT image. Here we can see that the non-credible region is at the posterior end of the bladder and that the segmentation has extended beyond the object boundary at that point.

A second example of this visualization is shown in Fig. 7. In the left pane we see the prostate surface with non-credible regions signaled by dark shading. In the right pane we see the intersection of this surface with a sagittal slice of the CT image. We recommend manual validation and correction of the prostate segmentation in the non-credible region.

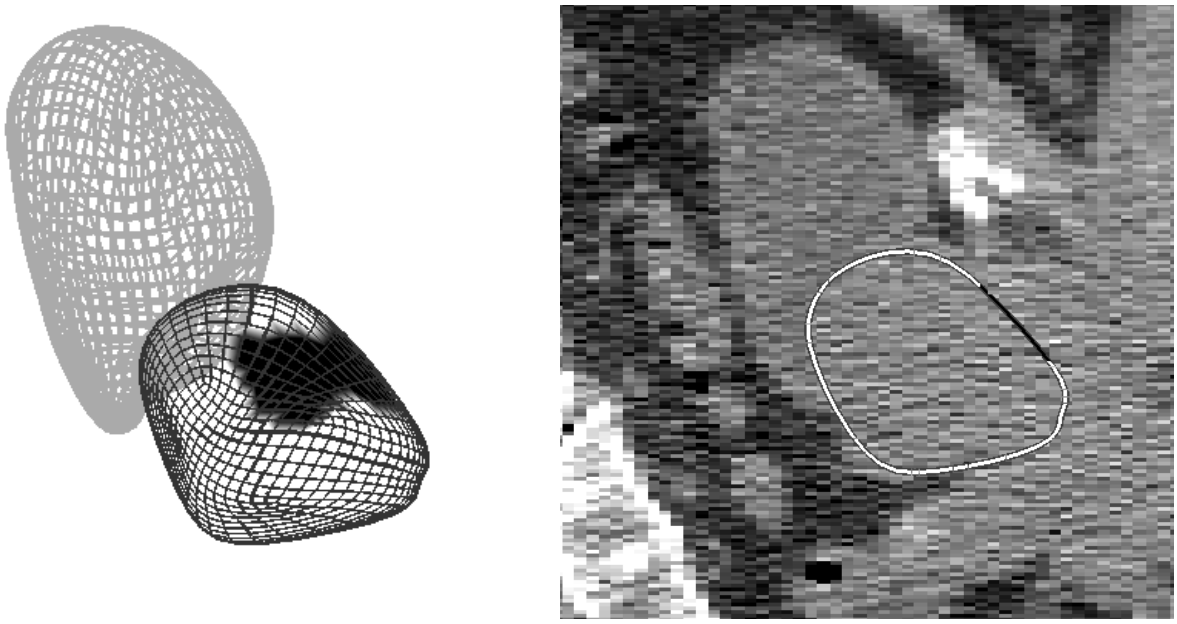

Figure 7. (Left) A display indicating a non-credible region (dark tiles) on a prostate segmentation surface. For orientation purposes, the bladder segmentation is shown with unshaded tiles. (Right) The intersection of the prostate surface with a sagittal slice of the CT image. Non-credible regions are indicated by the dark contour.

6. CONCLUSIONS

We have presented an automated method for signaling the location of non-credible regions on a segmentation surface. This method is based on detecting statistical outliers of a local image match function. We have applied this method to m-rep based segmentations of the bladder and prostate in 80 CT images. We validated these results with ROC analysis relating the local image match value with the distance from a representative point for each region to the a manual segmentation. The AUC for this ROC was 0.89 for the bladder and 0.92 for the prostate, indicating that our test was successful at distinguishing between those regions which are acceptably segmented and those that are grossly erroneous.

This validation relies on three simplifying assumptions that warrant further work. First, we allow the manual segmentation of each image to serve as the ground truth for each image. Given what is known about the intra- and inter- expert segmentation variability, a better estimate of the ground truth could be made by combining multiple manual segmentations as with the STAPLE[8] algorithm. An improved ground truth estimate would allow for a better measurement of localized segmentation errors. The second simplifying assumption is related to

how we measure localized segmentation errors. Currently we measure a single distance from the representative point for a region to the manual segmentation. An improvement would be to incorporate the distribution of distances from points within the region. The third simplifying assumption that we make is that the image match at neighboring regions is independent. Given that images are spatially correlated, this assumption is not likely to hold.

There are two significant consequences to the correlation of image data that our future work will address. On the one hand, we will modify our validation strategy to recognize neighbor relationships. New rules for labeling a region will be developed that account for the status of nearby regions. For example, when adjacent regions both have a large image match but only one has a significant error distance, it might be appropriate to consider both regions as true-positives. On the other hand, the research that lead to the RIQF image match is being extended to study conditional distributions of intensity quantiles. In the same situation described above it might be the case that when the RIQF for the acceptably segmented region is conditioned on its outlier neighbor, that conditional match falls below the threshold of the non-credibility test.

ACKNOWLEDGMENTS

This work was supported by NIH grant P01 EB02779.

REFERENCES

1. R. Broadhurst, J. Stough, S. Pizer, and E. Chaney, "A statistical appearance model based on intensity quantiles histograms," *ISBI* , 2006.
2. T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham, "Active shape models-their training and application," *Computer Vision and Image Understanding* **61**, pp. 38–59, January 1995.
3. S. Pizer, T. Fletcher, Y. Fridman, D. Fritsch, A. Gash, J. Glotzer, S. Joshi, A. Thall, G. Tracton, P. Yushkevich, and E. Chaney, "Deformable m-reps for 3d medical image segmentation," *International Journal of Computer Vision - Special UNC-MIDAG issue* **55**, pp. 85–106, November-December 2003.
4. L. Grady and G. Funka-Lea, "An energy minimization approach to the data driven editing of presegmented images/volumes," in *MICCAI*, pp. 888–895, 2006.
5. W. J. Niessen, C. J. Bouma, K. L. Vincken, and M. A. Viergever, "Error metrics for quantitative evaluation of medical image segmentation.," in *Theoretical Foundations of Computer Vision*, R. Klette, H. S. Stiehl, M. A. Viergever, and K. L. Vincken, eds., pp. 275–284, Kluwer, 1998.
6. S. Pizer, P. Fletcher, S. Joshi, A. Gash, J. Stough, A. Thall, G. Tracton, and E. Chaney, "A method & software for segmentation of anatomic object ensembles by deformable m-reps," *Medical Physics* **32**, pp. 1335–1345, May 2005.
7. P. T. Fletcher, C. Lu, S. M. Pizer, and S. Joshi, "Principal geodesic analysis for the study of nonlinear statistics of shape," *Medical Imaging, IEEE Transactions on* **23**, pp. 995–1005, 2004.
8. S. K. Warfield, K. H. Zou, and W. M. W. III, "Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation," *IEEE Trans. Med. Imaging* **23**(7), pp. 903–921, 2004.